Cloud Capacity Management

**1.Background in the area of study.**

**Definition-**

The objective of capacity planning is to specify the overall capacity level of cloud resources, so you can more effectively improve service performance, increase agility, and reduce cloud costs.

**Why does cloud need capacity management-**

1. **Provide insight into long-term IT planning.**

For example, capacity management can help determine which workloads to move to the cloud. Workloads with fast-changing capacities are ideal candidates for the cloud, where resource allocations can be easily scaled up and down.

**2-Determine which infrastructural and application architectures align with your needs.**

For instance, if you have a virtual server with routinely fluctuating capacity demands, you might find that serverless functions would be a better way to host that workload. Serverless functions allow you to allocate large amounts of resources for short periods in a more cost-effective and easy-to-manage way than is possible with virtual servers.

**3-Arrange the right people and tools.**

This is a step beyond your team knowing how many resources to allocate to workloads. It's important to find out if you have the organizational resources necessary to assign those resources. You'll need staff on hand to perform the necessary provisioning, and those workers should have the requisite skills to work with the tools you use to manage resource allocation.

**4-Avoid disruptions to users.**

Wrong-sized workloads can create problems for the people who expect a specific application to be ready for them when they need it. When your workload capacities are well managed, you minimize your risk of having applications or servers fail.

2.Significance

**Cost reduction**:

Strategic cloud capacity planning helps IT anticipate and plan for potential business changes that can impact cloud resource management. When IT understands business priorities and plans, they can control, track, and adjust resource capacity, consumption, and any related budgets or quotas accordingly, which can all have an impact on your bottom line.

**Performance improvement**:

Poor application performance can lead to poor user experiences, and even customer churn. IT can help avoid this by uncovering and fixing the performance bottlenecks from systems and applications as part of strategic cloud capacity planning. Additionally, it enables IT to make more informed decisions on how to achieve optimal performance in a cost-effective way.

**Flexibility and agility**:

Access to historical data and usage patterns - as part of effective cloud capacity planning - enables IT to effectively plan for unforeseen spikes in demand without disrupting the business.

PUNEET WILL CONTINUE

**3. How does cloud manage capacity**

Slide- Steps to manage cloud capacity

**1. Assess baseline capacity requirements**

First, determine how many cloud servers, application instances, databases and so on your team requires on average to maintain adequate performance. You'll need to know how many CPU, memory and storage resources each workload requires -- these are your baseline capacity requirements. It's important to remember that you shouldn't use that baseline to make resource allocations, especially if demands placed on the workloads often fluctuate. Still, knowing your baseline provides a starting point for capacity planning.

**2. Assess scalability needs**

Once you know the baseline requirements for each workload that you run in the cloud, examine the scalability they'll require. Evaluate how much variation occurs to workload demand between different times of day, days of the week or seasons of the year. Some of your cloud workloads will have higher scalability requirements than others. For instance, a website with a globally dispersed user base probably won't see as much fluctuation in usage in a full day as a website that caters to users in a specific geographic location, which likely will see most demand during that locale's daytime hours. Likewise, a website for a meal-delivery service will probably experience higher load during mealtimes than at other times of day.

**3. Make initial resource allocations**

For workloads that don't already run in the cloud, you'll need to set initial resource allocations before you start them. Plan to allocate 20% more resources to each workload than the baseline requirements dictate. This provides a healthy buffer in case demand unexpectedly jumps but doesn't unreasonably overprovision your environment.

**4. Set up autoscaling policies**

Mainstream public cloud providers allow you to create so-called autoscaling policies. With these policies in place, the cloud platform automatically increases or decreases the resource allocations assigned to your workloads based on the traffic thresholds you configure in the policies. You can apply autoscaling policies to most types of cloud workloads, including virtual machine instances, databases, containers and serverless functions. However, certain niche categories of cloud workloads, such as IoT devices, typically can't be managed using autoscaling.

**5. Collect and analyze capacity data**

Whether or not you configure autoscaling for your workloads, it's important to constantly assess how well the allocations work and adjust accordingly. Consider these metrics and factors:

**Slide 2**

**Cloud capacity management tools**

**1-Monitoring and log management.**

Data collected by monitoring and logging tools such as AWS CloudWatch, Azure Monitor and third-party monitoring platforms can help you keep track of performance trends and alert you to changing capacity needs.

Infrastructure as code. Infrastructure-as-code tools automate infrastructure setup and resource allocation, so it is much easier and faster to reconfigure allocations in response to capacity changes.

**2-Cost calculators.**

To manage the financial aspects of capacity planning, the cost-prediction tools that cloud providers offer are useful. They can help evaluate the costs associated with different resource allocations or workload types.

**3-Rightsizing and cost management.**

Cloud providers offer tools designed to help predict capacity requirements. AWS has a cost management tool, as does Microsoft Azure. Some third-party application performance management (APM) tools also offer right-sizing features.

4. **Future trends**

1- Investigating more efficient techniques for optimization

Text, letter

Description automatically generated

For understanding future trends

<http://www.diva-portal.org/smash/get/diva2:707752/fulltext01.pdf>